Case Studies and Regression Analysis

Jason Seawright

j-seawright@northwestern.edu

August 11, 2010

Examples of Strengths and Weaknesses

Case Selection

Case Selection

1. Study the entire population.

Case Selection

1. Study the entire population.
2. Take a random sample.

Case Selection

1. Study the entire population.
2. Take a random sample.
3. Follow some rule for deliberate case selection.

Mill's Methods

Mill's Methods

- Method of Agreement

Mill's Methods

- Method of Agreement
- Method of Difference

Mill's Methods

- Method of Agreement
- Method of Difference
- Etc..

Crucial Cases

Crucial Cases

- Theory assigns high likelihood to an outcome in a particular case that, for (all, or most, or a major) competing theory has low likelihood.

Regression Analysis

- In a survey of 1000 articles in 10 leading political science journals, 49% used statistics (Bennett, Barth, and Rutherford 2003).
  - Presumably, most of them involve some variant of regression.

Regression Analysis

- In a survey of 1000 articles in 10 leading political science journals, 49% used statistics (Bennett, Barth, and Rutherford 2003).
  - Presumably, most of them involve some variant of regression.
- A search in JStor for the word "regression" finds 10,404 relevant articles.

Regression Analysis

- Almost no matter what you work on, you will have to interact with regression-based studies.

Choosing Cases

- Case-selection rules:

Choosing Cases

- Case-selection rules:
  - Random sampling

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases

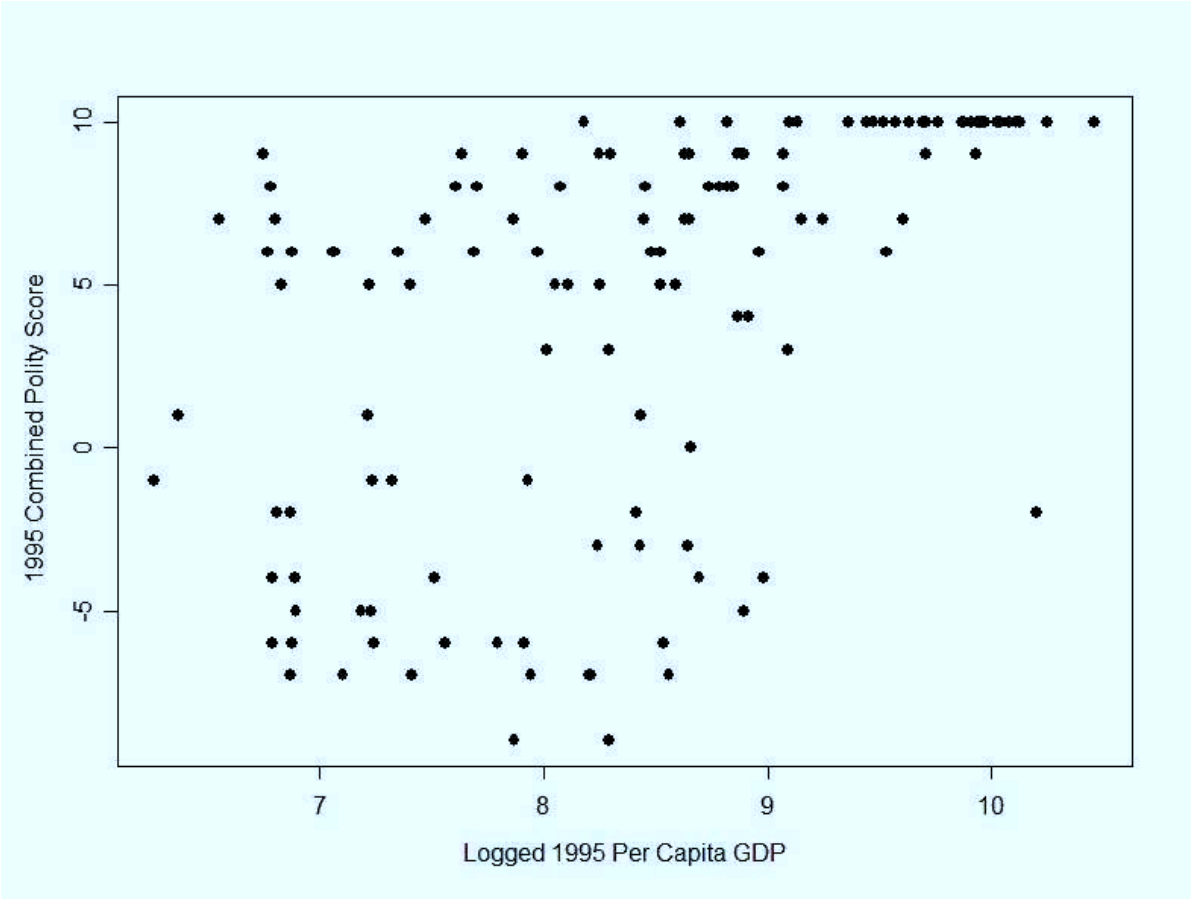
Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases
  - Influential cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases
  - Influential cases
  - Most-similar cases

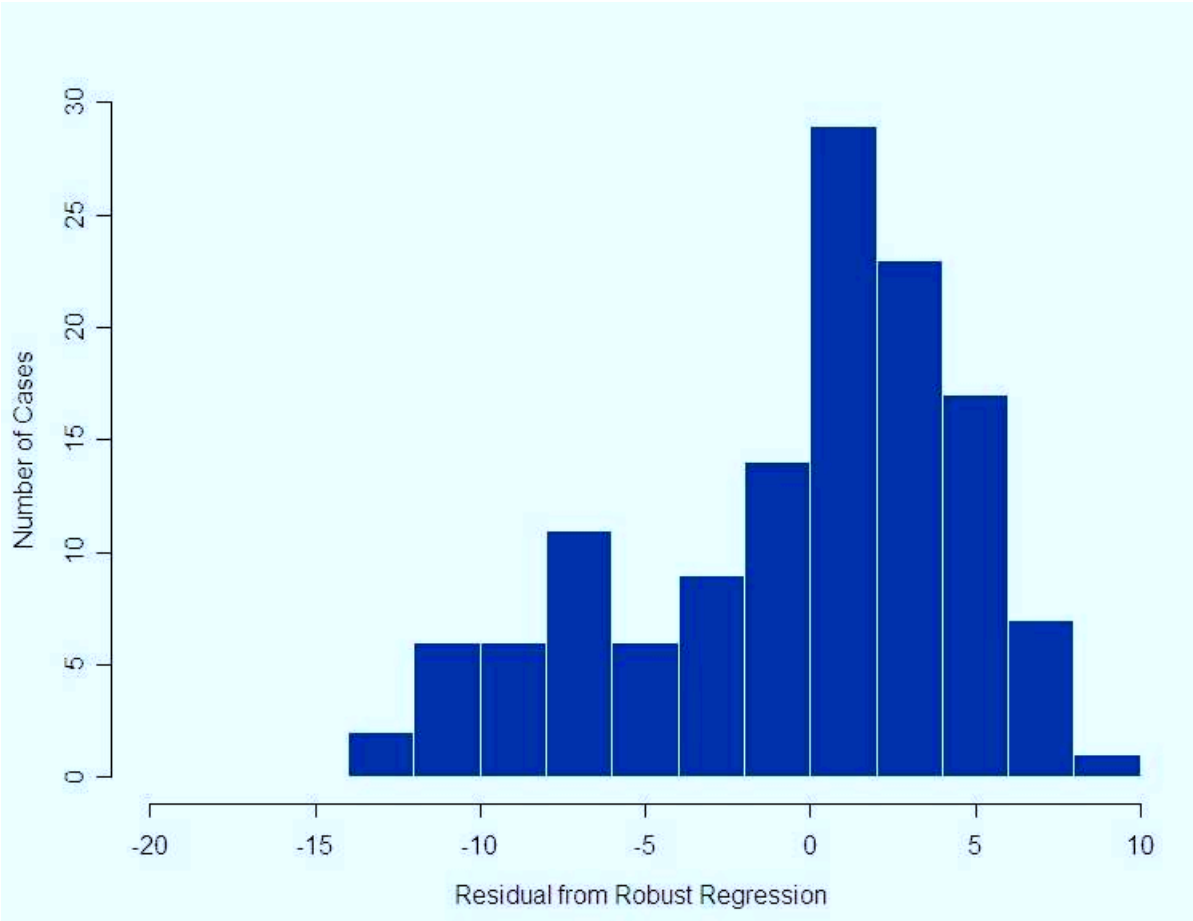
Running Example



Typical Cases

$$\text{Typicality}_i = -\text{abs}[y_i - \text{E}(y_i|x_{1,i}, x_{2,i}, \ldots, x_{k,i})] \quad (1)$$

Typical Cases



Extreme Cases

$$\text{Extremity}_i = |\frac{x_i - \bar{x}}{s}| \tag{2}$$

Deviant Cases

$$\text{Deviantness}_i = -\text{Typicality}_i \tag{3}$$

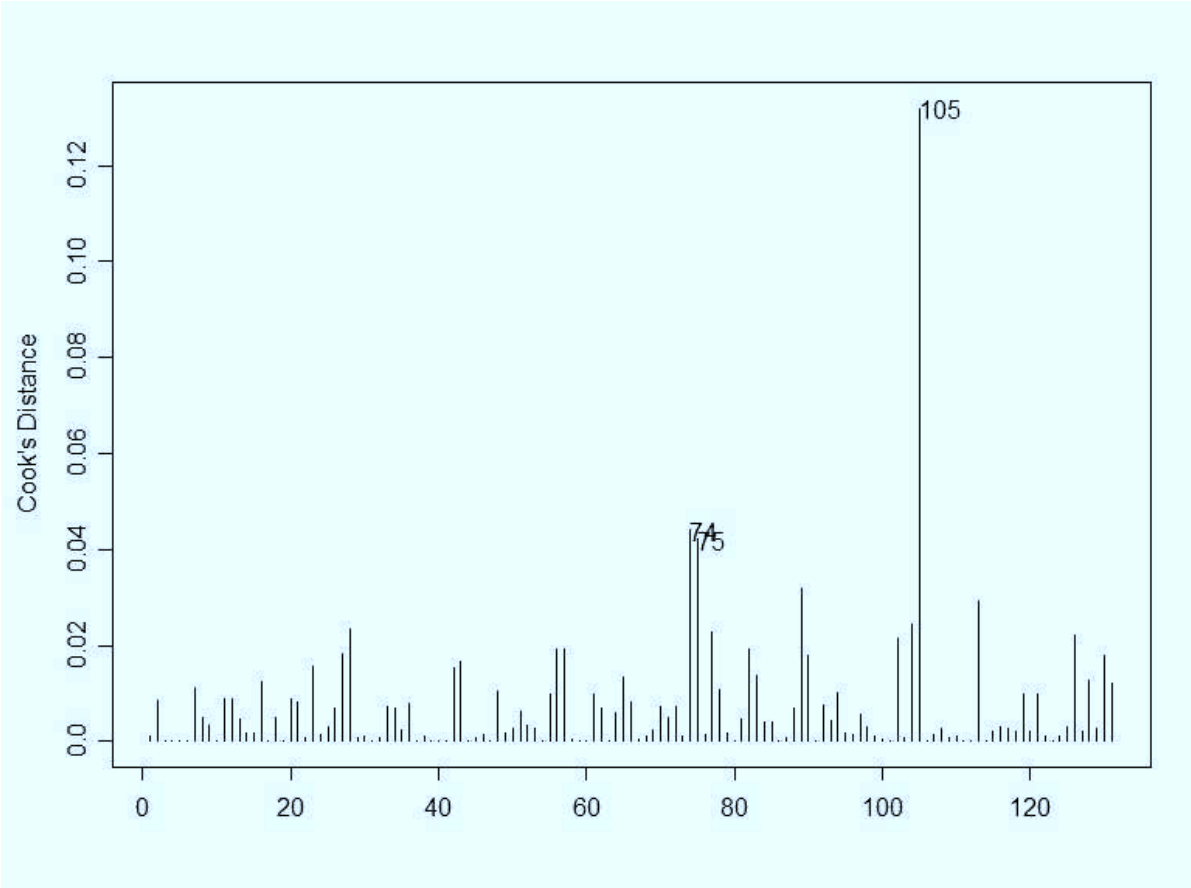
Influential Cases

- Cook's distance is a statistical measure of how much the overall regression result would change if a given case is deleted.

Influential Cases

- Cook's distance is a statistical measure of how much the overall regression result would change if a given case is deleted.
- A Cook's distance score of 1 or more usually is regarded as representing substantial influence.

Influential Cases



Most-Similar Cases

- Matching techniques are an automated way of finding most similar cases.

Measurement Error in Y

$$Y_i^* = Y_i + \delta_{Y,i}$$

Measurement Error in Y

$$Y_i^* = Y_i + \delta_{Y,i}$$

Random Sampling

Measurement Error in Y

Typical/Deviant Cases:

$$e_i = Y_i - \mathbb{H}_{i,.} Y + \delta_{Y,i}$$

Measurement Error in Y

Influential Cases Strategy:
Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

Measurement Error in Y

Extreme Cases:

$$Y_i^* = Y_i + \delta_{Y,i}$$

Measurement Error in Y

Most-Similar Cases

Measurement Error in Y

Most-Similar Cases

Most-Different Cases

Measurement Error in X

$$X_i^* = X_i + \delta_{X,i}$$

Measurement Error in X

$$X_i^* = X_i + \delta_{X,i}$$

Random Sampling

Measurement Error in X

Typical/Deviant Cases:

$$e_i = Y_i - X_i\hat{\beta}^* - \delta_{X,i}\hat{\beta}^*$$

Measurement Error in X

Influential Cases Strategy:
Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

Measurement Error in X

Extreme Cases:

$$X_i^* = X_i + \delta_{X,i}$$

Measurement Error in X

Most-Similar Cases

Measurement Error in X

Most-Similar Cases

Most-Different Cases

Omitted Variables

$$e_i = d_i + \gamma \tilde{Z}_i, \text{ where } \tilde{Z}_i = Z_I - \mathrm{E}(Z_i|X_i)$$

Omitted Variables

$e_i = d_i + \gamma \tilde{Z}_i$, where $\tilde{Z}_i = Z_I - \mathrm{E}(Z_i|X_i)$

Random Sampling

Omitted Variables

Typical/Deviant Cases:

$e_i = d_i + \gamma \tilde{Z}_i$

Omitted Variables

Influential Cases Strategy:
Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

Omitted Variables

Extreme Cases:

Omitted Variables

Extreme Cases:

For confounders, extreme on X may be a good strategy.

Omitted Variables

Extreme Cases:

For confounders, extreme on X may be a good strategy.

Extreme on Y maximizes:

$\hat{Y}_i + d_i + \gamma \tilde{Z}_i$

Omitted Variables

Most-Similar Cases

Omitted Variables

Most-Similar Cases

Most-Different Cases

Pathway Variables

$$W_i = \nu + \mu X_i + \omega_i$$
$$Y_i = \alpha + \tau W_i + \sigma_i$$

Pathway Variables

$$W_i = \nu + \mu X_i + \omega_i$$
$$Y_i = \alpha + \tau W_i + \sigma_i$$

Random Sampling

Pathway Variables

Typical/Deviant Cases:

$e_i = \tau\omega_i + \sigma_i$

Pathway Variables

Influential Cases Strategy:
Maximizes the product of the error term and the weighted average distance of the right-hand-side variables from their means.

Pathway Variables

Extreme Cases:

Pathway Variables

Extreme Cases:

$$W_i = \nu + \mu X_i + \omega_i$$

Pathway Variables

Extreme Cases:

$W_i = \nu + \mu X_i + \omega_i$

Extreme on Y maximizes:

$Y_i = \alpha + \tau W_i + \sigma_i$

Pathway Variables

Most-Similar Cases

Pathway Variables

Most-Similar Cases

Most-Different Cases

Summary: Analytic Arguments

	Deviant	Influential	Ext. X	Ext. Y
Error in Y	Good	Mixed	Poor	Good
Error in X	Mixed	Mixed	Good	Poor
Confound	Good	Mixed	Mixed	Good
Pathway	Good	Mixed	Good	Mixed

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.
- Add measurement error, omitted variables, etc.

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.
- Add measurement error, omitted variables, etc.
- 2 SD Rule

Simulation Results

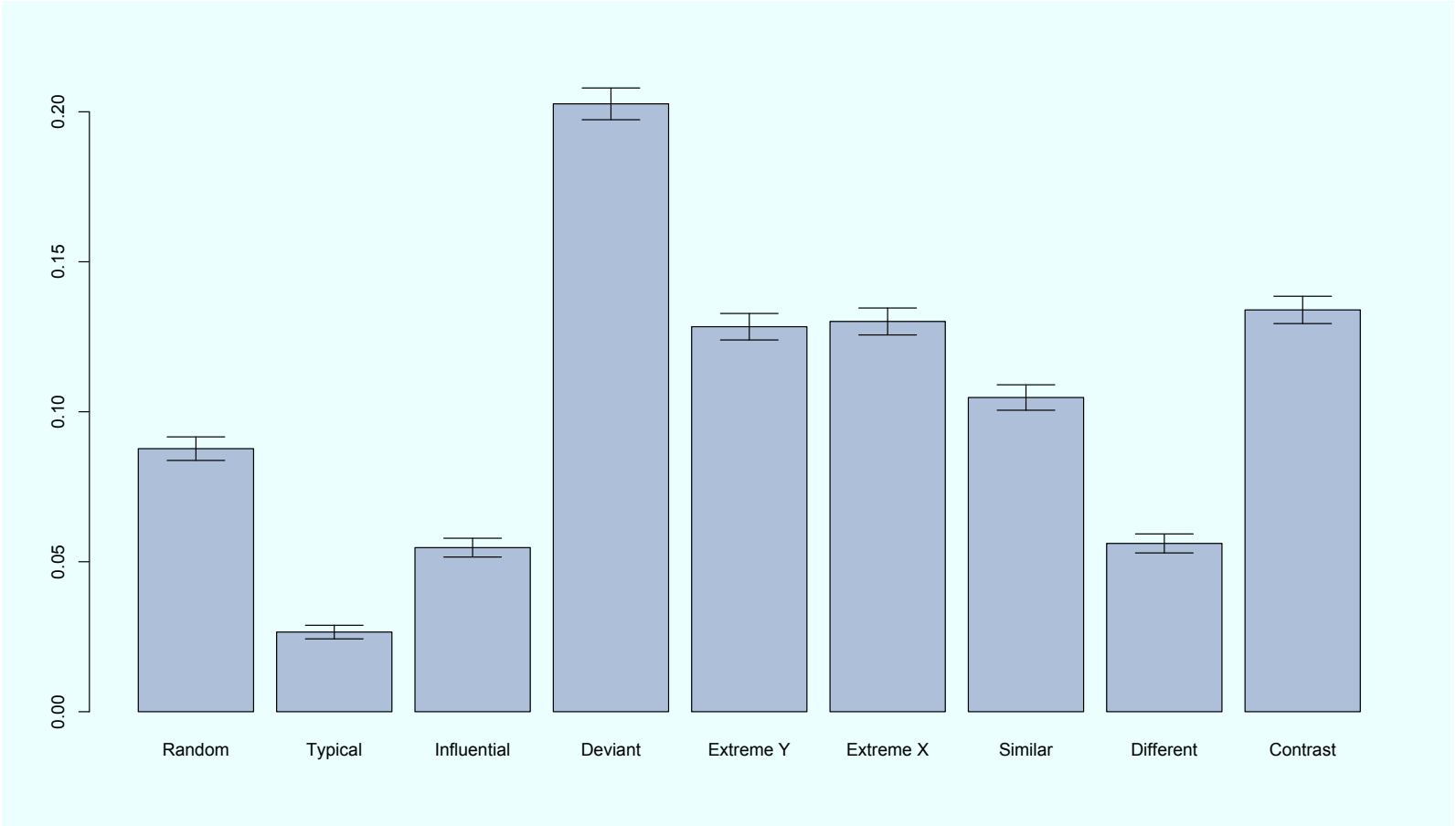


Figure: Case Selection for Finding Confounder.

Simulation Results

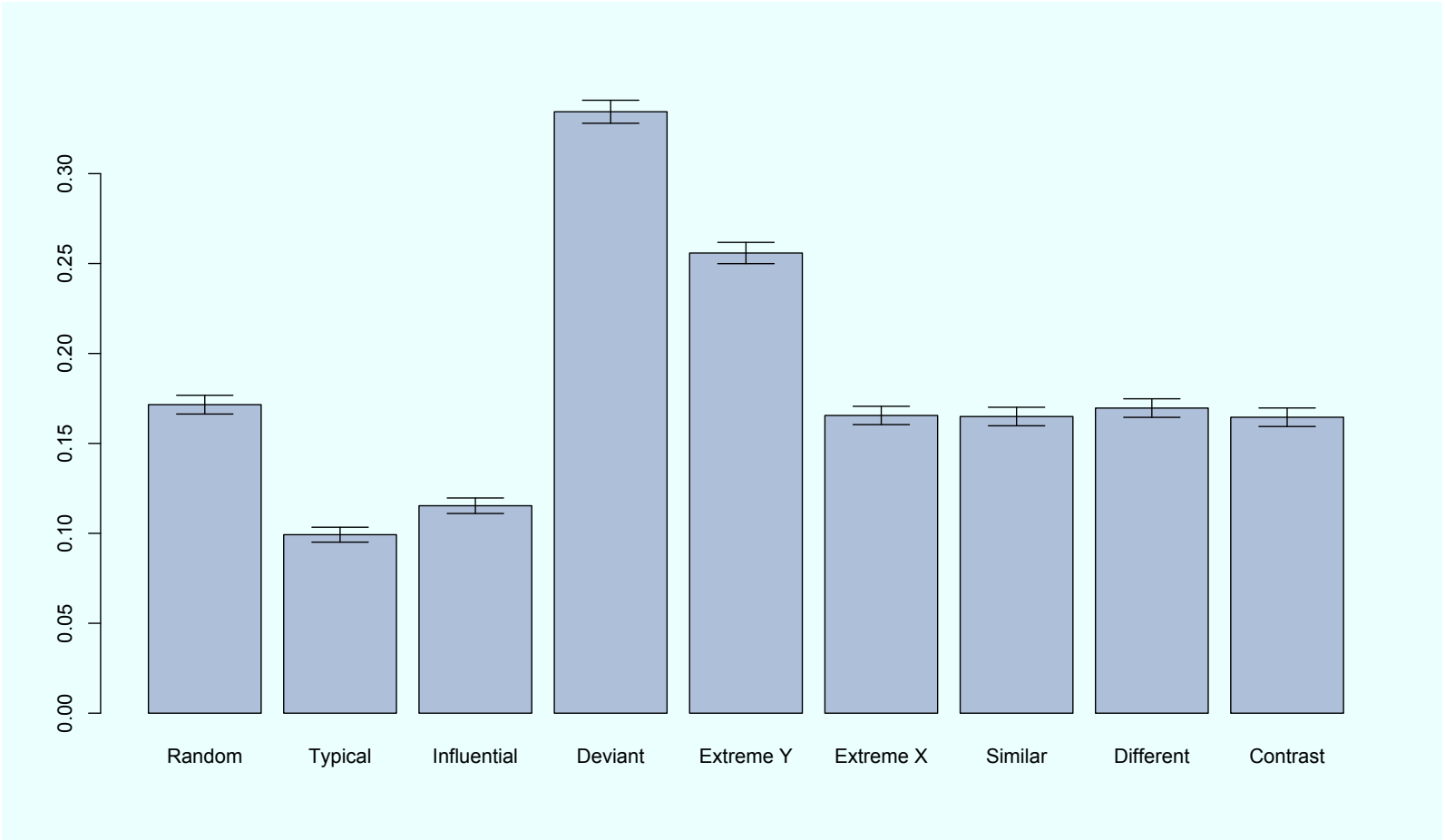


Figure: Case Selection for Other Causes.

Simulation Results

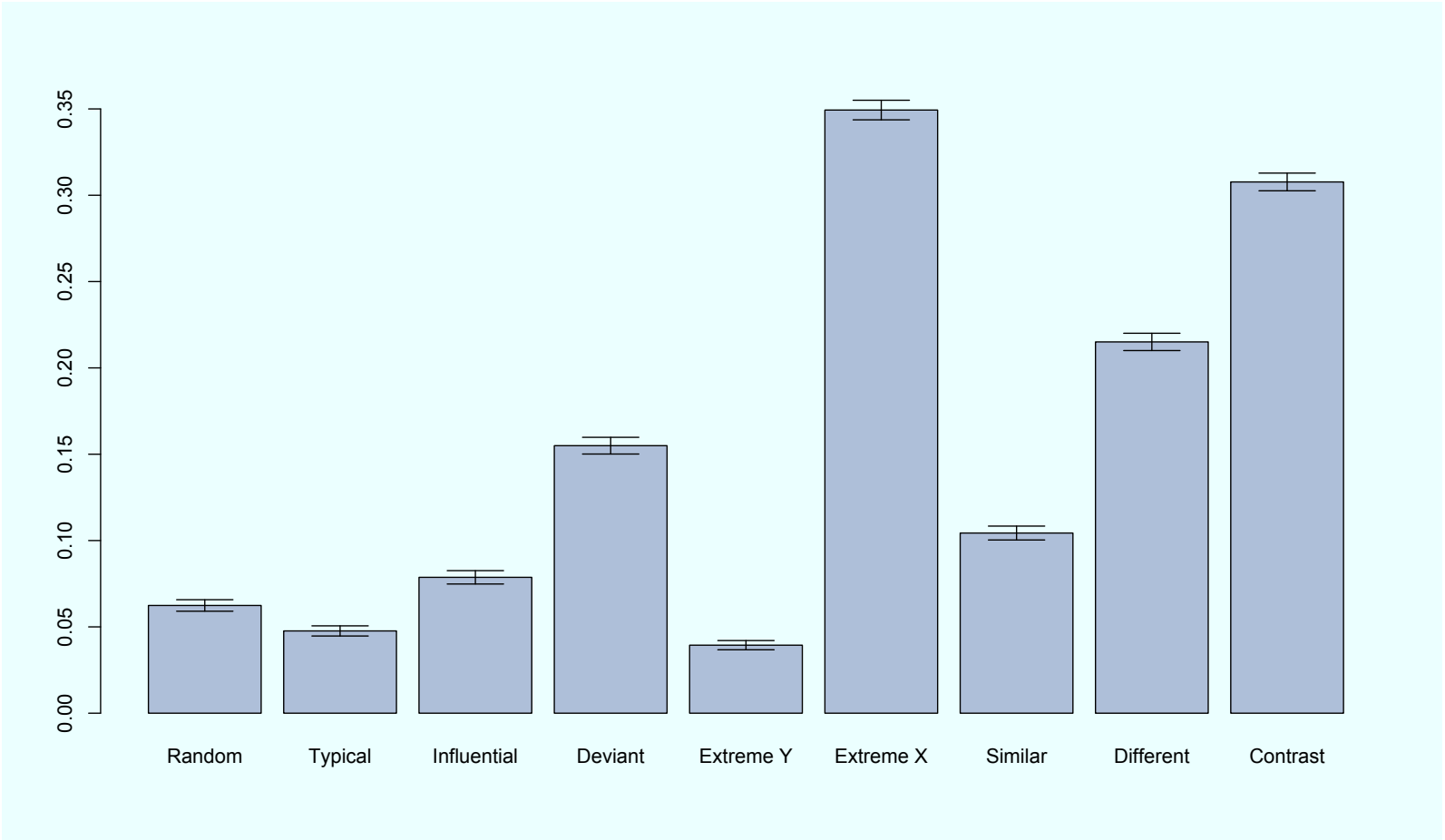


Figure: Case Selection for Exploring Mechanisms.

Simulation Results

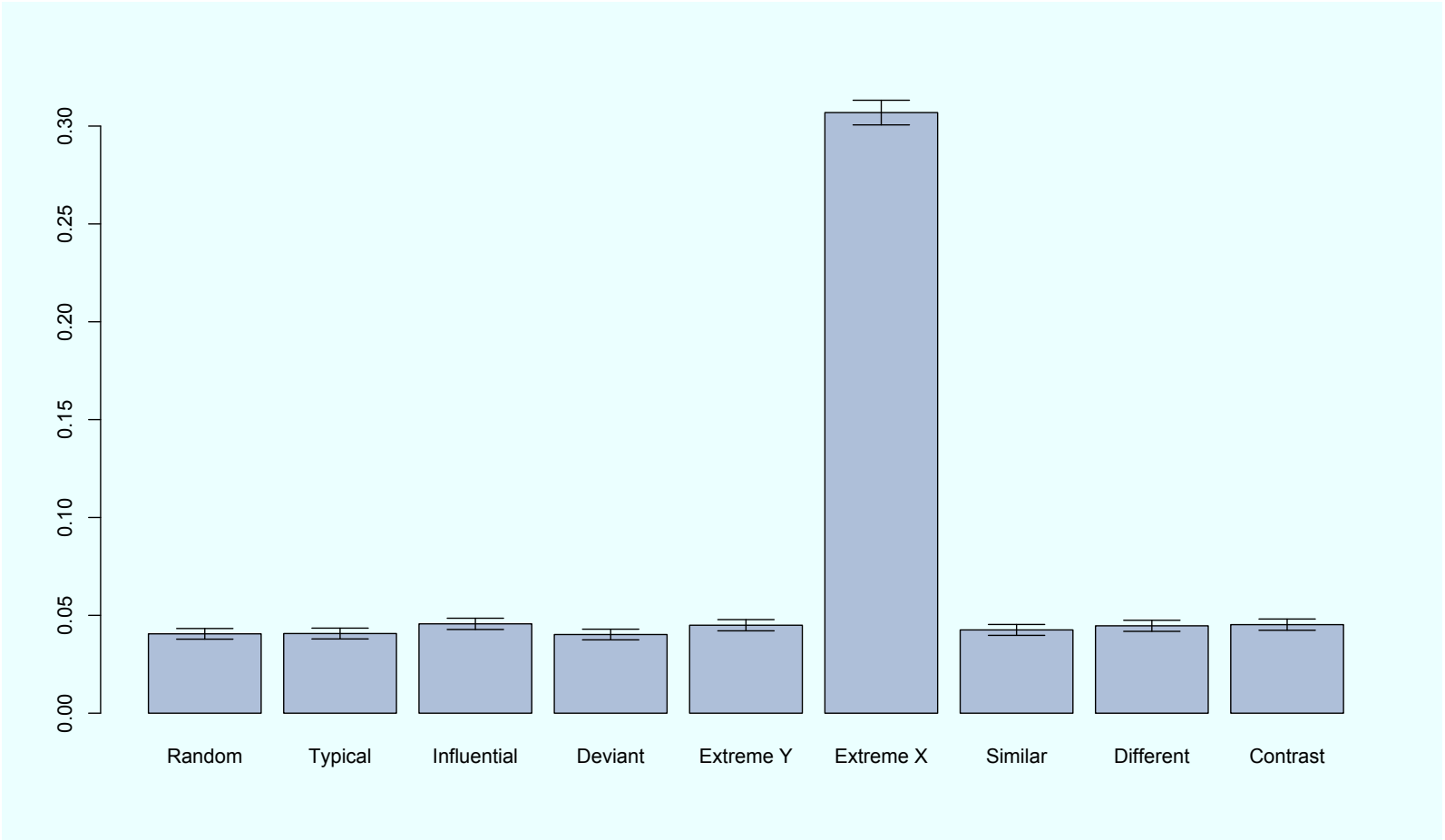


Figure: Case Selection for Error in X.

Simulation Results

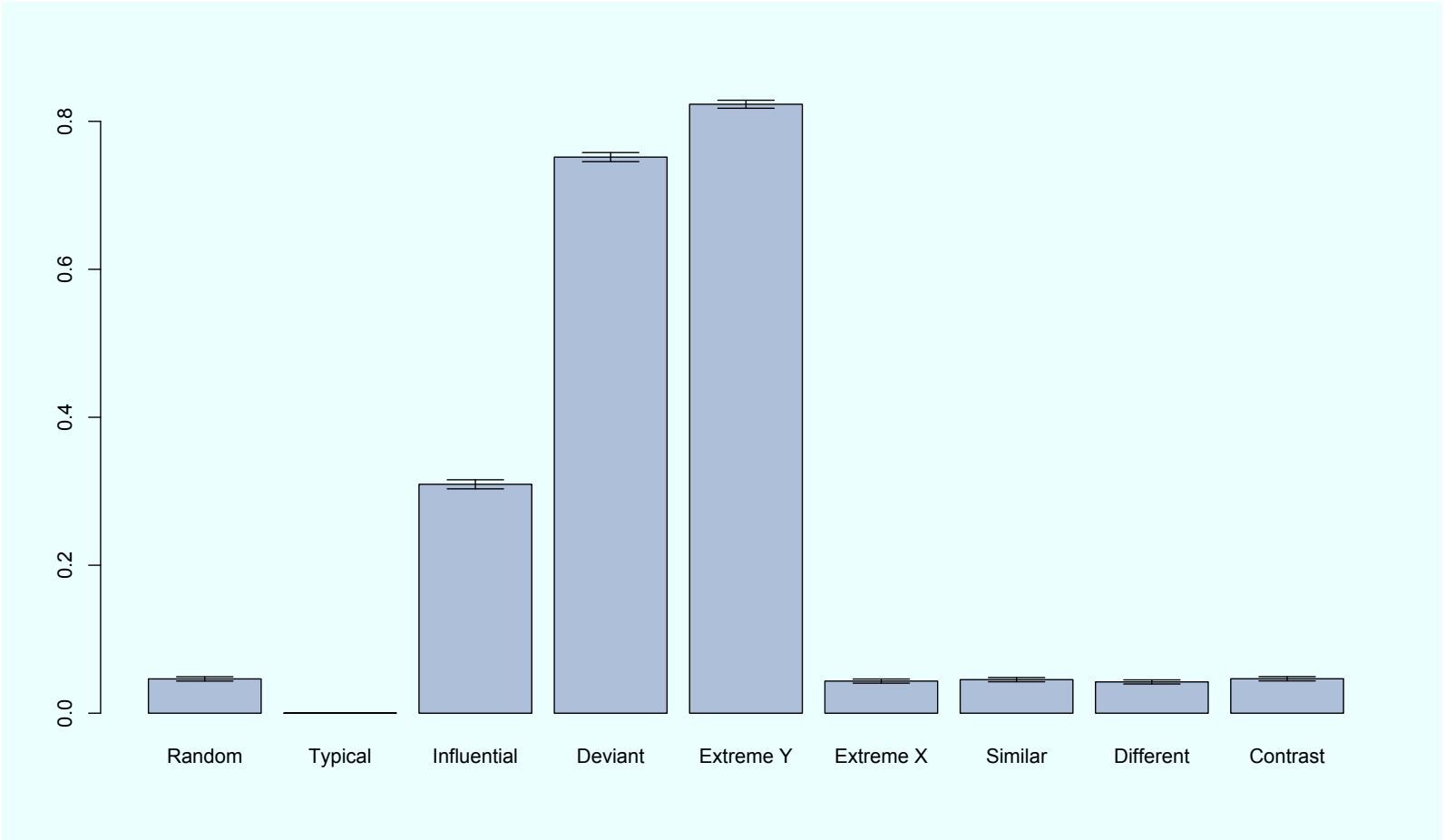


Figure: Case Selection for Error in Y.

Simulation Results

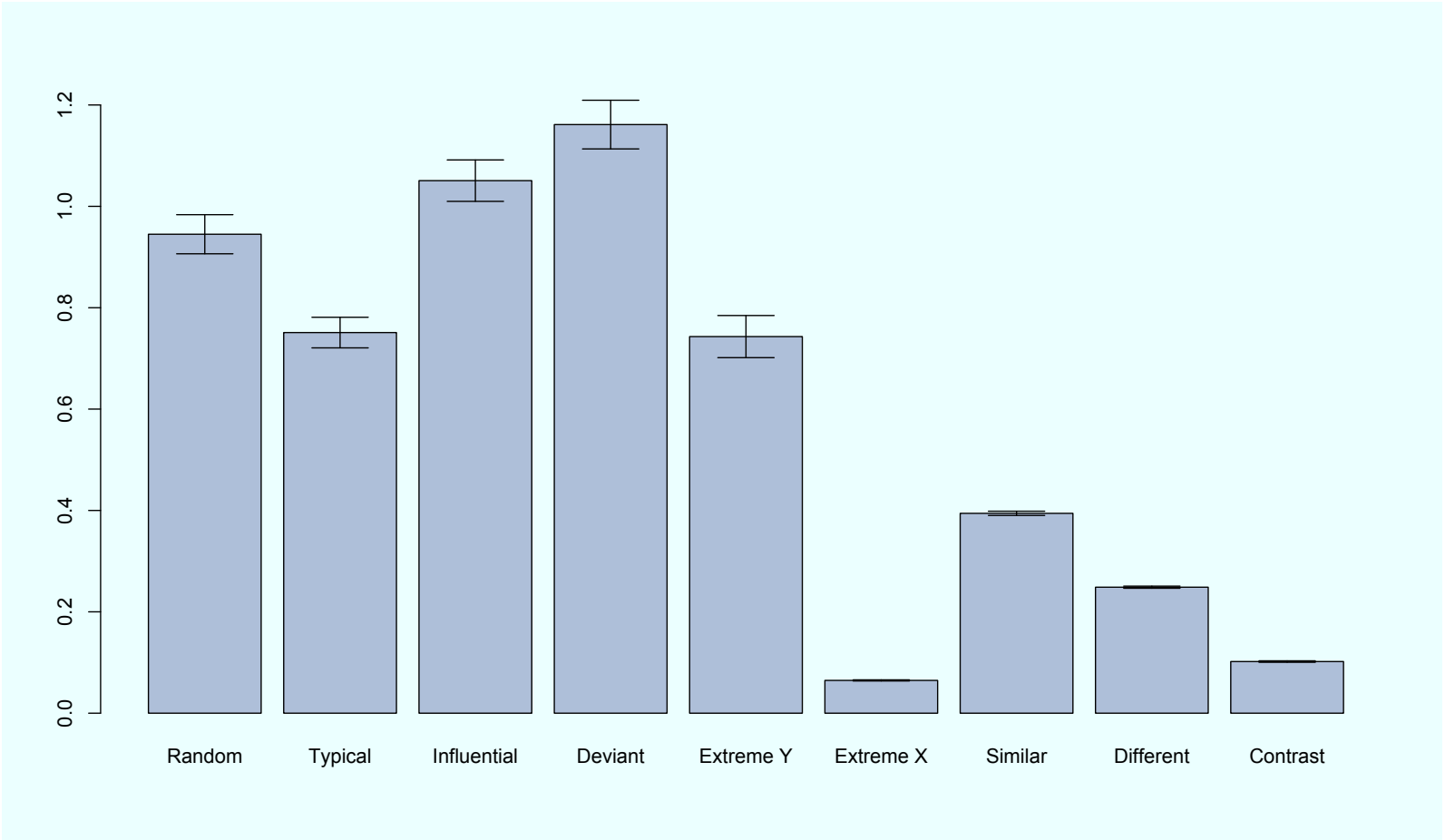


Figure: Case Selection for Estimating Overall Slope.

Case-selection software in R

Assignment

Implement each case-selection technique for a data set of interest, or off my website. Be prepared to discuss what kinds of cases you get, and whether they seem on first glance to be useful, tomorrow.